**Report**:

**Approach**:

* Used datahub to ingest metadata, and monitoring the queries and schema
* Used “great expectations“ as the data quality framework
* Data was fetched from the API and pushed to Snowflake, could have done it through other sources like CSV or Postgresql, but either Snowflake or Great Expectations did not have either of the 2 connectors available, plus both platforms work seamlessly with Snowflake as a data source, therefore I made the decision to use Snowflake
* Used python to set up the ingestion pipeline
* Used great expectations commands to create data sources, suites for data quality checks and checkpoints for keeping track of the quality checks
* Since DataHub support for Great Expectations has expired, and the Validations Tab was not displayed, I opted out to using Great Expectations reports. Using the docs build command
* The data quality checks could have been many, but i used the following
* All columns must conform to their data types
* VARCHAR columns were checked by using regex
* Max population values should not be greater than the total population of US which is 333 Million
* No null values for any of the columns, as all of them are mandatory
* Unique checks could not be implemented, the generated ID column was set to be a primary key, therefore no need to check for uniqueness
* A Virtual Environment was used within VS Code so as to isolate the development from the local environment.

**Challenges**:

* DataHub does not support Validations anymore, therefore a substitute was needed, for that I decided upon Great Expectations, as it implements data quality checks seamlessly and easily, plus it is extremely good for reporting, and can be installed over the cloud to make it available for many users, just like DataHub
* When setting up Great Expectations I kept encountering the problem that Great Expectations was not initializing, but then as a last resort, I created a venv environment, and then installed all the packages therein, and then after starting a new Terminal within VS Code, it ran successfully. As before, it was running on my local machine as a package rather than a module.

**Suggestions for Future**:

* DataHub Cloud should be setup over Kubernetes
* A new action to Great Expectations YAML that triggers a python scripts that adds failure description to datahub scema column called "Test"
* Since there were only about 900 rows, the robustness of the data quality rules could not be verified.